ChinaXiv合作期刊

知识管理论坛 ISSN 2095-5472 CN11-6036/C
Knowledge Management Forum

E-mail: kmf@mail.las.ac.cn
http://www.kmf.ac.cn

chinaXiv:202310.03153v1

【学术探索】

基于大规模文本数据情感挖掘的企业舆情研究

◎ 吴联仁[1] 李瑾颉[2] 齐佳音[1]

[1] 上海对外经贸大学工商管理学院 上海 201620 [2] 北京邮电大学经济管理学院 北京 100876

摘要：[目的 / 意义] 大数据环境下，文本挖掘和情感分析技术在产品、服务等网络点评分析中得到越来越广泛的应用。通过对大规模文本数据情感挖掘，研究影响企业舆情的关键要素。[方法 / 过程] 基于中国大陆 292 个城市 103 878 家酒店的 2 500 多万条网络点评数据，挖掘企业在线舆情，识别影响顾客服务体验的关键内容要素。采用探索性因子分析方法对关键要素进行归类，并通过多元回归分析得出评论内容要素与顾客总体满意度之间的关系。[结果 / 结论] 酒店客房要素和电器要素对酒店业顾客总体满意度影响最大。本研究方法和结论为服务企业营销和管理的大数据商业分析研究提供参考。

关键词：网络点评 文本挖掘 情感分析 企业舆情 商业分析

分类号：C93

引用格式：吴联仁 , 李瑾颉 , 齐佳音 . 基于大规模文本数据情感挖掘的企业舆情研究 [J/OL]. 知识管理论坛 , 2016, 1(6): 457-463[引用日期]. http://www.kmf.ac.cn/p/1/79/.

1 引言

在过去的数年中，信息技术在社会、经济、生活等各个领域不断渗透和推陈出新。在移动计算、物联网、云计算等一系列新兴技术的支持下，社交媒体、协同创造、虚拟服务等新型应用模式持续拓展着人类创造和利用信息的范围和形式。基于信息和网络的生产模式创新正在将人类社会带入"第三次工业革命"时代。新兴信息技术与应用模式的涌现，使得全球数据量呈现出前所未有的爆发式增长态势。预计到 2020 年，全球被创建和被复制的数据总量将达到 35ZB。与此同时，数据的多样性、低价值密度、实时性等复杂特征日益显著。冯芷艳等[1]指出大数据背景下，商务管理研究也面临着前所未有的挑战。

大数据时代，随着电子商务网站、社区型网站和第三方评论网站的发展以及在旅游、酒店行业的普及应用，网络上出现了大量的顾客对酒店的点评内容。截至 2014 年底，从全国各大中文网站能够采集到的酒店顾客点评数量已达到千万级。这些点评内容实际上是顾客在网

基金项目：本文系国家自然科学基金重点项目"面向不确定性的 Web2.0 用户创作内容管理研究"（项目编号：71231002）研究成果之一。

作者简介：吴联仁 (ORCID: 0000-0001-7886-6494)，讲师，博士，E-mail: lianrenwu@suibe.edu.cn；李瑾颉 (ORCID: 0000-0002-9269-7268)，博士研究生；齐佳音（ORCID：0000-0001-7162-4898），教授，博士生导师。

收稿日期：2016-08-06 发表日期：2016-12-29 本文责任编辑：王传清



知识管理论坛
2016 年 第 6 期（总第 6 期）

chinaXiv:202310.03153v1

络环境下对酒店所提供产品与服务的自发的"问卷调查"结果，是顾客在享受酒店产品和服务后对酒店满意度的详细描述。对这些点评进行有效的采集和分析，将能够代替传统的问卷调查评价，并且能够弥补传统问卷样品有限性和问题局限性的不足。

伴随着大数据时代的到来和自然语言处理技术的快速发展，文本挖掘（text mining）方法——对具有丰富语义的文本进行分析从而理解其所包含的内容和意义的过程——逐渐被认为是更可靠和经常使用的研究方法。在管理科学研究中，文本挖掘方法经常被用来处理网络点评等非结构化数据。如黄敏学等 [2] 和李杰等 [3] 采用文本挖掘方法研究了网络环境下的网络口碑或点评。在旅游和酒店业，文本挖掘方法也渐渐开始被应用 [4]。目前，大部分的酒店网络点评内容研究主要是对内容特征属性、评论内容分词的统计分析和聚类分析。例如，L. Zhou 等 [5] 对评论提及酒店各要素的数量进行了统计，给出了各要素的占比。Z. Xiang 等 [6] 采用文本分析方法研究顾客体验与顾客满意度间的关系。熊伟 [7] 对点评中提及酒店各要素的数量进行了统计，并计算了评论在各要素上的评价得分，做了各项服务体验要素与总体评价的相关分析。

随着大数据文本挖掘研究的深入，情感分析（sentiment analysis），又称意见挖掘 (opining mining) 开始应用到网络点评这种非结构化的自然语言处理中 [8]。张紫琼等 [9] 指出文本情感分析是指通过语义分析技术对文本的主客观性、观点、情绪、极性的挖掘和分析，对文本的情感倾向做出分类判断。E.Cambria[10] 表示基础的文本情感分析是对文本情感极性分析和文本情感极性强度分析。杨立公等 [11] 将情感极性分为两极，即正面（positive）的赞赏和肯定、负面（negative）的批评与否定。也有学者在正面和负面之间加入了中性（neural），如 H. Li 等 [12] 首先通过词频分析方法，对评论中各因子出现的频数进行统计，其次采用聚类分析对出现的因子进行聚类，最后是统计了各因子的正面、中性和负面点评的占比。另外一些学者采用情感极性强度分析网络点评，如丁于思等 [13] 将顾客满意度分为很不满意、不满意、一般、满意和很满意 5 个等级。情感分析在大数据环境下对企业顾客洞察、市场营销策略和商业模式创新起到了重要作用。如李实等 [14] 挖掘中文网络客户评论的产品特征及情感倾向。刘羽等 [15] 在李实等基础上，进行观点挖掘的产品特征提取。

2 数据采集与处理

本研究使用的数据集由北京众荟信息技术有限公司（http://www.jointwisdom.cn/）数据应用事业部提供。众荟信息是目前国内旅游、酒店行业主要的大数据挖掘与应用服务提供商。数据集包括了 2 500 多万条网络点评，涉及国内 292 个城市的 103 878 家酒店。数据来源于国内 8 个主流中文点评网站，分别为到到网、大众点评网、艺龙、美团、阳光旅行、住哪儿、去哪儿和携程。数据收集时间窗口为 2014 年 1 月 1 日 -2014 年 12 月 31 日。

借助众荟信息的自然语言处理和语义分析技术，对酒店网络点评进行酒店特征词的抽取和情感分析。作者基于众荟信息的酒店网络点评数据处理结果，提炼出 80 多个影响酒店顾客服务体验的特征词，构成了本研究的特征词集合。分别统计特征词关注度（attention），即特征词在顾客网络点评中被顾客提及的频次，特征词的参与度 (engagement)，即特征词的关注度与酒店数的比率，特征词的满意度（satisfaction），即特征词正面提及的频次占总频次的比例（具体计算方法见第 3 部分情感分析模型）。表 1 给出特征词关注度排名前 30 的特征词。

从表 1 可以看出，最受顾客关注的是位置，这与丁于思等的研究结果一致。另外关于位置的参与度也是最高的，为 32.99，即每家酒店顾客网络点评中平均提及位置的频次为 32.99。在满意度方面，满意度最高的是娱乐，其次是酒吧和交通，都超过了 90%。而满意度

知识管理论坛，2016(6):457–463
DOI：10.13266/j.issn.2095–5472.2016.053

chinaXiv:202310.03153v1

排在倒数三位的是隔音、异味和电梯，分别为 14.06%、17.49% 和 18.11%，均未超过 20%。这三个酒店顾客体验要素是酒店经营管理者应该重点关注的。

表 1 酒店顾客点评中排名前 30 的特征词

特征词	关注度	参与度	满意度	特征词	关注度	参与度	满意度
位置	3 426 992	32.99	89.79%	停车场	210 518	2.03	71.85%
服务	2 812 568	27.08	88.01%	总机	164 169	1.58	71.49%
交通	2 309 254	22.23	92.22%	空调	162 454	1.56	28.75%
价格	2 260 528	21.76	80.00%	电器	147 212	1.42	44.47%
环境	1 624 840	15.64	84.73%	热水	140 618	1.35	49.58%
餐厅	1 420 470	13.67	63.06%	酒吧	120 060	1.16	94.62%
卫生	1 382 394	13.31	85.76%	异味	113 012	1.09	17.49%
设施	1 239 264	11.93	65.39%	洁具	109 824	1.06	35.24%
房间	1 086 660	10.46	88.19%	游泳池	92 326	0.89	65.97%
隔音	479 592	4.62	14.06%	电视	79 396	0.76	28.67%
前台	379 400	3.65	60.40%	大堂	67 476	0.65	57.74%
网络	373 076	3.59	66.24%	礼宾	66 876	0.64	78.64%
装饰	372 162	3.58	51.46%	娱乐	55 186	0.53	98.35%
床	344 730	3.32	67.08%	电梯	43 002	0.41	18.11%
卫生间	333 986	3.22	40.53%	家具	40 158	0.39	47.82%

3 情感分析模型

每条网络点评都是顾客对酒店设施及服务的真实反馈，但是这种非结构的文字并不利于科学的数据分析。笔者基于情感分析技术，将用自然语言描述的用户点评，转化为结构化的用户情感数据库，点评文本挖掘与情感分析流程见图 1。其中，顾客在点评中所表达的对酒店软硬件某一方面的看法及情感态度，可以理解为该顾客在点评中对酒店该要素进行了一次满意程度的“投票”，并可以被转化为顾客对酒店该方面的情感表达。顾客的情感分为正向和负向。具体的点评分析样例见表 2。

首先对酒店特征词在评价集 { 正向，负向 } 上的频次进行统计。得到酒店特征词的情感频次向量 $F(W_i)=\{F(W_i)^+, F(W_i)^-\}$，其中 (i=1,2,3,…, 30)，$F(W_i)^+$ 为特征词的正面观点频次，$F(W_i)^-$ 为特征词的负面观点频次。因此，酒店特征 W_i 的满意度为：

$$S(W_i)=\frac{F(W_i)^+}{F(W_i)^+ + F(W_i)^-} \quad (1)$$

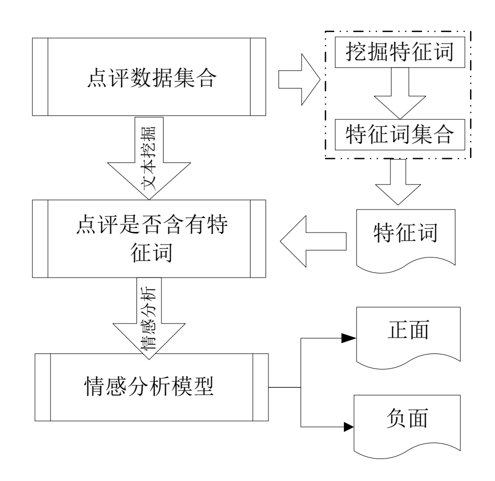


图 1 点评文本挖掘与情感分析流程



chinaXiv:202310.03153v1

表 2 网络点评分析样例

点评数据	酒店特征	情感词
地理位置也好，很多景点就在附近，服务让人觉得很棒。 大堂门面比较 low。 酒店地理位置好，酒店环境不错，服务态度很好！	位置 服务 环境 大堂	好 很棒、很好 不错 low

↓

分析结果		
特征词	情感极性	频次
位置	正面	2
服务	正面	2
环境	正面	1
大堂	负面	1

本研究以城市为单位，城市酒店顾客总体满意度 $CityS_i(i=1,2,3,\ldots,292)$，即为城市酒店顾客点评中正向情感频次占城市总情感频次的比例。

4 统计分析

4.1 探索性因子分析

对酒店顾客点评中抽取的排名前 30 的特征词，利用 SPSS 进行探索性因子分析。Bartlett 检验结果（P=.000）说明各变量间具有相关性。KMO 统计量为 0.772，在 0.7 以上，可以进行探索性因子分析。图 2 为因子分析的碎石图。本研究提取了 6 个公因子，累计方差贡献率为 58.53%。

表 3 是进行方差最大旋转后的因子载荷矩阵。表 3 中给出了载荷大于 0.5 的因子，并将载入的 20 个特征词分为 6 类。第 1 类是电器，包括电视、网络、空调和电器；第 2 类是客房，包括卫生间、装饰、床和房间；第 3 类是位置，包括位置、环境和交通；第 4 类是娱乐，包括娱乐、游泳池和酒吧；第 5 类是服务，包括礼宾、服务和前台；第 6 类是卫生，包括卫生和异味。

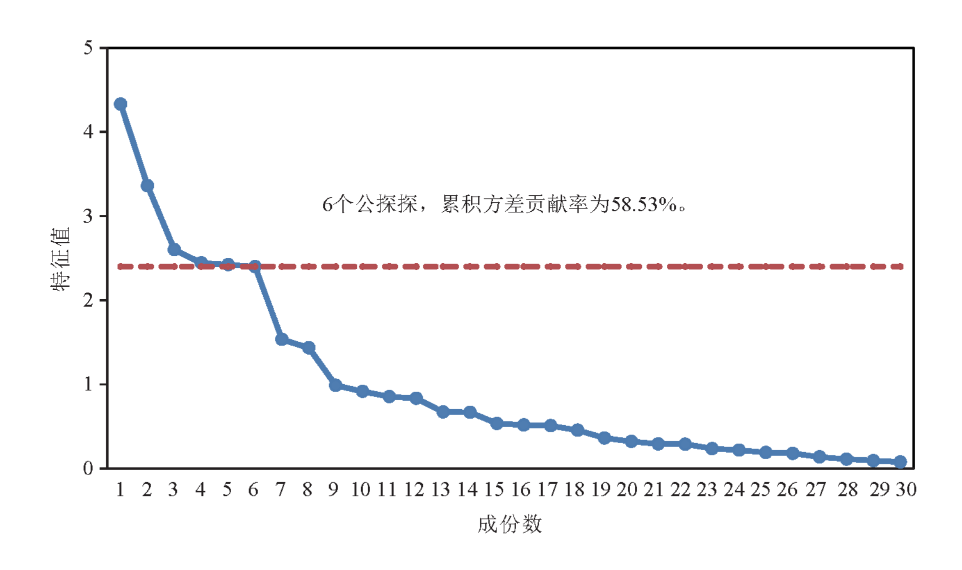


图 2 碎石图分析

记提取的公因子为 $U_i(i=1,2,3,4,5,6)$，U_i 的情感频次 F 为公因子所包括的特征词情感频次之和。

公因子 U_i 的满意度为

$$S=\sum \beta_{ij} S\left(W_{ij}\right) \tag{2}$$

知识管理论坛，2016(6):457–463
DOI：10.13266/j.issn.2095–5472.2016.053

chinaXiv:202310.03153v1

表 3 探索性因子分析结果 (只显示因子载荷大于 0.50)

提取的公因子	特征词	特征值	累积方差贡献率	因子载荷
电器	电视	4.331	14.436%	0.814
	网络			0.802
	空调			0.719
	热水			0.664
	电器			0.634
客房	卫生间	3.361	25.538%	0.816
	装饰			0.710
	床			0.684
	客房			0.515
位置	位置	2.601	34.308%	0.874
	环境			0.826
	交通			0.619
娱乐	娱乐	2.444	42.454%	0.813
	游泳池			0.800
	酒吧			0.610
服务	礼宾	2.423	50.530%	0.738
	服务			0.709
	前台			0.502
卫生	卫生	2.399	58.528%	0.729
	异味			0.639

式中：$S(W_{ij})$ 为公因子 U_i 下第 j 个特征词的满意度；β_{ij} 为公因子 U_i 下第 j 个特征词在公因子 U_i 中的权重。

以公因子位置为例，其情感频次为 F=F(位 置)+F(交 通)+F(环 境)=3 426 992+ 2 309 254+1 624 840=7 361 086

公因子位置的满意度为

$$S=\frac{F(位置)}{F}S(位置)+\frac{F(交通)}{F}S(交通)+\frac{F(环境)}{F}S(环境)$$

=3 426 992/7 361 086×89.79%+2 309 254/7 361 086×92.22%+1 624 840/7 361 086×84.73%=89.44（数据参考表 1）

基于上述方法，以城市为单位，分别计算每个公因子的满意度。

4.2 多元回归分析

将 292 个城市酒店总体满意度值作为因变量，城市酒店 6 个公因子满意度作为自变量进行多元线性回归，回归结果如表 4 所示：

表 4 多元线性回归分析结果

模型	非标准化系数		标准系数	T	Sig.
	B	标准误差	Beta		
常量	.232	.019		12.341	.000
客房	.185	.010	.448	18.861	.000
卫生	.114	.017	.181	6.692	.000
电器	.168	.010	.312	16.546	.000
娱乐	.058	.012	.134	4.893	.000
位置	.166	.019	.166	8.797	.000
服务	.091	.015	.164	5.882	.000

注：因变量：城市总体满意度；调整 R^2: 0.968

表 4 的结果显示，在显著性水平 p=0.01 下，6 个因子的系数都是显著的。并且客房和电器两个因子的标准化系数最大，分别为 0.448 和 0.312。这说明，客房和电器对酒店顾客满意度的影响很大。客房因子主要包括卫生间、装饰、床和客房 4 个二级因子，电器因子主要包括电视、网络、空调、热水和电器 5 个二级因子。这 9 个因子可以被认为是酒店提供的核心产品。目前酒店作为一个提供住宿功能的场所，如果满足了顾客的基本需求，即提高顾客在客房因子和电器因子的满意度，将会提升酒店顾客的总体满意度。

其次是卫生因子，标准化系数为 0.181，也对酒店顾客总体满意度产生较为重要的影响。卫生因子包括卫生和异味 2 个二级因子。在酒店提供核心产品保障了顾客的基本需求的基础上，如酒店需要进一步提高顾客的总体满意度，应着重在卫生因子上提高顾客的满意度。

系数最低的是娱乐因子，为 0.134。娱乐因子包括娱乐、游泳池和酒吧 3 个二级因子。在 6



个因子中，娱乐对酒店顾客总体满意度的影响最低。这可能是因为娱乐作为一项增值服务，对顾客来说，不是顾客的必需产品。因此，顾客娱乐因子满意度的提高对顾客总体满意度的提升影响不大。

5 总结与讨论

随着电子商务网站、社区型网站和第三方评论网站的发展，中国酒店业也迎来了大数据时代。虽然，在许多学科中大数据分析已经被描述为一个新的研究范式。然而作者发现，在旅游和酒店服务业领域充分和深入发掘数据分析功能的研究还较少。本研究采用文本挖掘和情感分析的方法，归类大量的酒店顾客网络点评，评估这些数据的质量，分析酒店顾客体验要素与顾客总体满意度之间的影响关系。这项研究的创新之处在于其数据规模，有别于传统调查研究在数据量上的局限。本研究只是在酒店大数据分析中的初步探索，但已经取得了一些实质性的结论，希望为酒店等服务企业开展营销和管理的商务分析研究提供一些借鉴。

参考文献：

[1] 冯芷艳 , 郭迅华 , 曾大军 , 等 . 大数据背景下商务管理研究若干前沿课题 [J]. 管理科学学报 , 2013, 16(1): 1-9.

[2] 黄敏学 , 王峰 , 谢亭亭 . 口碑传播研究综述及其在网络环境下的研究初探 [J]. 管理学报 , 2010, 7(1): 138-146.

[3] 李杰 , 张向前 , 陈维军 , 等 . C2C 电子商务服装产品客户评论要素及其对满意度的影响 [J]. 管理学报 , 2014, 11(2): 261-266.

[4] 丁于思 , 肖轶楠 . 基于网络点评的五星级酒店顾客满意度测评研究 [J]. 经济地理 , 2014(5): 182-186.

[5] Zhou L, Ye S, Pearce P L, et al. Refreshing hotel satisfaction studies by reconfiguring customer review data [J]. International journal of hospitality management, 2014, 38: 1-10.

[6] Xiang Z, Schwartz Z, Gerdes J H, et al. What can big data and text analytics tell us about hotel guest experience and satisfaction?[J]. International journal of hospitality management, 2015, 44: 120-130.

[7] 熊伟 , 高阳 , 吴必虎 . 中外国际高星级连锁酒店服务质量对比研究 —— 基于网络评价的内容分析 [J]. 经济地理 , 2012, 32(2): 160-165.

[8] 周立柱 , 贺宇凯 , 王建勇 . 情感分析研究综述 [J]. 计算机应用 , 2008, 28(11): 2725-2728.

[9] 张紫琼 , 叶强 , 李一军 . 互联网商品评论情感分析研究综述 [J]. 管理科学学报 , 2010, 13(6): 84-96.

[10] Cambria E, Schuller B, Xia Y, et al. New avenues in opinion mining and sentiment analysis[J]. IEEE intelligent systems, 2013 (2): 15-21.

[11] 杨立公，朱俭，汤世平 . 文本情感分析综述 [J]. 计算机应用 , 2013, 33(6): 1574-1607.

[12] Li H, Ye Q, Law R. Determinants of customer satisfaction in the hotel industry: an application of online review analysis[J]. Asia Pacific journal of tourism research, 2013, 18(7): 784-802.

[13] 丁于思 , 肖轶楠 . 五星级酒店服务质量评价指标体系研究 —— 基于网络点评内容分析 [J]. 消费经济 , 2014, 30(3): 64-69.

[14] 李实 , 叶强 , 李一军 , 等 . 挖掘中文网络客户评论的产品特征及情感倾向 [J]. 计算机应用研究 , 2010, 27(8): 3016-3019.

[15] 刘羽 , 曹瑞娟 . 基于观点挖掘的产品特征提取 [J]. 计算机应用与软件 , 2014, 31(1): 81-84.

作者贡献说明：
吴联仁：负责文献调研及论文起草；
李瑾颉：负责数据处理、统计分析和计量模型构建；
齐佳音：负责论文设计及研究指导。

chinaXiv:202310.03153v1

知识管理论坛，2016(6):457–463
DOI：10.13266/j.issn.2095–5472.2016.053

Research on Enterprise Public Opinions Based on Large-scale Text Data Sentiment Mining

Wu Lianren[1] Li Jinjie[2] Qi Jiayin[1]
[1]School of Management, Shanghai University of International Business and Economics, Shanghai 201620
[2] School of Economics and Management, Beijing University of Posts and Telecommunications, Beijing 100876

Abstract: [Purpose/significance] In the era of big data, text mining and sentiment analysis technologies have been widely used in the analysis of online reviews (ORs). Through the large-scale text data mining, the key factors influencing the public opinion of enterprises are studied. **[Method/process]** We collected more than twenty-five million hotel online reviews from 103 878 hotels, identifying key content elements that affected the customer service experience. **[Result/conclusion]** Through the exploratory factor analysis and the multiple regression analysis, the authors explore the relationships between the hotel customer experience and satisfaction. It is hoped that this study sets an example for the development of business analytics in enterprises marketing and management.

Keywords: online review text mining sentiment analysis enterprise public opinion business analysis

chinaXiv:202310.03153v1